Saving Literary Digital Text (SLDT)

Sara Marhaba, Afnan Alomairi, Arwa Magdy, Hatoon Alazwari, Fatma heiba
Supervised by: Dr. Areej Althubaity

Project ID: UQU-CS-2021S-15 Computer Science Department, Umm Al-Qura University, Makkah, Saudi Arabia

Abstract

Arabic is the fourth spoken language on social media. Therefore, Natural Arabic Language Processing (NLP) has gained a lot of attention from the research community in the past few years. In this project, we extract a collection of digital Arabic literature published on Twitter and build a system that categorizes and classifies tweets mainly into hate and non-hate ones. based on Machine Learning model, namely Random Forest that has produced high accuracy results. The system will preserve these tweets and the identity of their authors, as well as other relevant data, such as date and time. in a database, then the system displays this information on our website.

Introduction

Middle Eastern's people are one of the largest populations on Twitter. In recent years, more Arabic writers, especially the youth ones have used Twitter to publish their literature work to a wider audience. Such a platform has provided them with a service to share their writings, but it lacks the ability to preserve their creative literature works for the long run. Hence, we propose a system named Saving Literary Digital Text (SLDT) where Arabic tweets will be scrapped, pre-processed, and classified into hate or non-hate tweets by applying machine learning algorithms such as Random Forest (RF), Our results have shown that the **RF algorithm produces 95.45% in accuracy**, 9**8.63% in precision**, **92.17% in recall**, and **95.29% in F1-score**. Classified tweets are displayed on a website named "ميراث الأدباء" which means Writers' Legacy in English language. We believe that our project will be significant and will add more value to native and non-native Arabic speakers as well as to the authors of these tweets.

Conclusion

In this study ,we propose a system named SLDT (Saving Literary Digital Text) for detecting hate speech in Arabic literary texts scraped from Twitter and displaying it on a website. We have used the four most common criteria, namely: accuracy, precision, recall and F1-score to evaluate and to compare the performance of various Machine Learning models. the Random Forest algorithm with oversampling technique has achieved the highest scores in each of the four metrics. To be precise, it has 95.45% in accuracy, 98.63% in precision, 92.17% in recall, and 95.29% in F1-score. As a result, we have decided to display the literary tweets that are classified by Random Forest model with over-sampled data on our website. We believe that SLDT will make Arabic literature thrive, since it is preserving modern forms of hate and non-hate and it creates more digital content in Arabic language.

Methodology

SLDT is composed of **five steps**, as follows:

1. **Scraping:** we scraped tweets from **97 Arabic writers'** accounts with some information such as: **the writer's profile username, and the writer's name.**

2. **Dataset Formation:** we create a dataset and used some of the existing datasets, in total we collected **3285 words**

3. **Data Preprocessing:**
   - **Data cleaning**
   - **Data normalization**
   - **Remove stop words:**
   - **Stemming**
   - **Annotation**

4. **Feature Extraction:** unigram and bigram TF-ID.

5. **Machine Learning Models:**
   - **Extra Tree Classifier**
   - **Random Forest**
   - **Gradient Boosting**
   - **Support Vector Machine**
   - **Naïve Bayes**

Technologies

Technologies and language we have use in SLDT:

Contact

Sara Marhaba:
saramarhaba@hotmail.com
Afnan Alomairi:
ayomairi@outlook.com
Arwa Magdy:
alaraberoro33@gmail.com
Hatoon Alazwari:
hatoone333@gmail.com
Fatma heiba:
fatmayahyahaiba@gmail.com
Dr. Areej Althubaity:
akkthubaity@uqu.edu.sa

Evaluation

Figure 1 illustrates the amount of hate and non-hate tweets. We have compared the performance of different Machine Learning models according to various evaluation metrics, namely accuracy, F1, precision, and recall.
We have non-balanced data Thus, we had tested the models on three scenarios: using non-balanced data, under sampling balanced data, and over sampling balanced data.

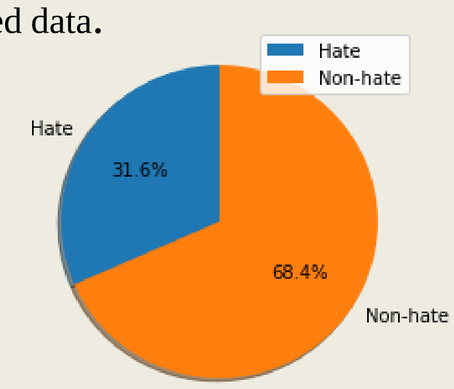

Figure 1: The ratio of hate and non-hate tweets in the data.

Figure 2 shows that Random Forest has highest scores in each of the four metrics, it has **95.45% in accuracy**, **98.63% in precision**, **92.17 %in recall**, and **95.29% in F1-score.**

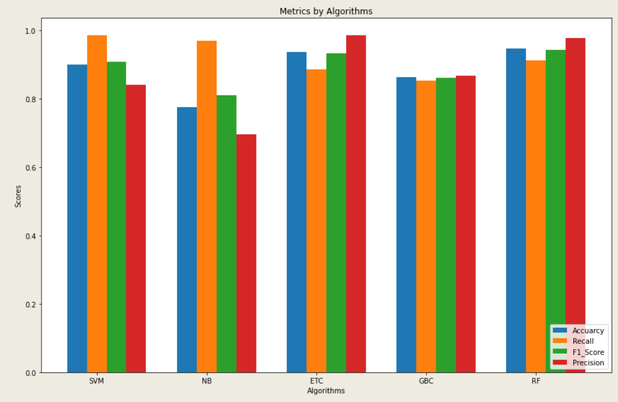

Figure 2: Over-sampling data

SLDT website

Figure 3 shows one of the pages on the SLDT website.

Figure 3: Text Page